Title: Performance Implications of Using Oracle XMLDB to Implement CDISC ODM for Clinical Study Informatics System

Authors:
Shaohua Alex Wang[a], Huey Cheung[a], Frank Pecjak[a], Barg Upender[a], Adam Frazin[a], Raj Lingam[b], Sarada Chinatala[a], Gladys Wang[b], Marc Kellogg[a], Robert L. Martino[a], Yang Fann[b], and Calvin Johnson[a]

[a]Center for Information Technology, National Institutes of Health, Bethesda, MD
[b]National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD

Technical Areas:
Web-based Delivery of Medical Information
Software Systems in Medicine
Medical Database & Information Systems

Author Information:
Shaohua Alex Wang
High Performance Computing and Informatics Office
Center for Information Technology
National Institutes of Health
12 South Drive
Building 12A, Room 2008
Bethesda, MD 20893-5624

Voice: 301-402-5895
Fax: 301-402-2867

Email: wangal@mail.nih.gov

Introduction

Promoting clinical research is a major priority in the new strategic plan for the National Institute of Neurological Disorders and Stroke (NINDS). The web-based Clinical Study Informatics System (CSIS) is a major component of an integrated Clinical Informatics and Management System (CIMS), which is being developed for NINDS intramural clinical researchers. In addition to CSIS, CIMS also contains the Protocol Tracking Management System (PTMS), which supports protocol submission, approval, and monitoring of the protocol review process; and a data integration module, which provides data warehousing services to collect data from a variety of data sources for analysis and potentially allowing extramural research through CSIS.

Database Design Considerations

CIMS is being developed as a web-based n-tiered architecture. For the third tier, the persistent data layer, we have chosen Oracle Corporation's Oracle 9i database with the XMLDB package installed.

User requirements dictate that the database be generic enough to allow investigators to create arbitrary clinical forms without the intervention of a programmer or database administrator. If a new entity or attribute is needed, the appropriate structure must be created automatically with all necessary relationship constraints and proper indexing to insure data integrity and optimal performance. The dynamic nature of such a system leads us to consider a metadata approach to data management. A metadata approach utilizes a general structure where only high-level relationships are defined. Specific information and relationships are maintained as row elements rather than column elements in the structure.

For the storage structure in a metadata approach, we consider two options: (1) relational *entity-attribute-value* (EAV); and (2) XML Schemas defined within Oracle XMLDB. The other user requirement is easy transport and representation of the data to external systems. To promote this information exchange among researchers and ease clinical trial data submission to FDA, we have considered implementing the standard Operational Data Model (ODM) created by Clinical Data Interchange Standards Consortium (CDISC), an open, multidisciplinary, non-profit organization committed to the development of industry standards [1]. The XML-based ODM model supports electronic acquisition, exchange, submission and archiving of clinical trial data and metadata for medical and biopharmaceutical product development. The model represents study metadata, study data and administrative data associated with a clinical trial. The model provides the ultimate in flexibility for representing study and form definition. This model lends itself directly to the metadata approach described above. An additional advantage of considering the CDISC model is our technology approach is not dictated by the model and can be evaluated on performance factors.

Representing the data structure with a strong metadata layer requires a common vocabulary be used to define new data elements. The CDISC standard addresses this with their Submission Data Standard (SDS) model. Where ODM is the structure of the metadata, SDS is the vocabulary of the metadata. Utilizing this flexible terminology, new terms and concepts can be defined and, if necessary, tied to existing terms through the use of industry dictionary standards (such as UMLS [2] or LOINC).

Figure 1 depicts the current data storage model design. Clinical data can be transported to and from the database as an XML file through FTP and HTTP protocols. The XML based CDISC model fits well with this transport mechanism. The database layer may also be accessed via SQL statements through direct JDBC or Oracle network services connections. Clinical data can be stored utilizing the Oracle XMLDB functions or in traditional relational tables, following an *entity-attribute-value* (EAV) format.

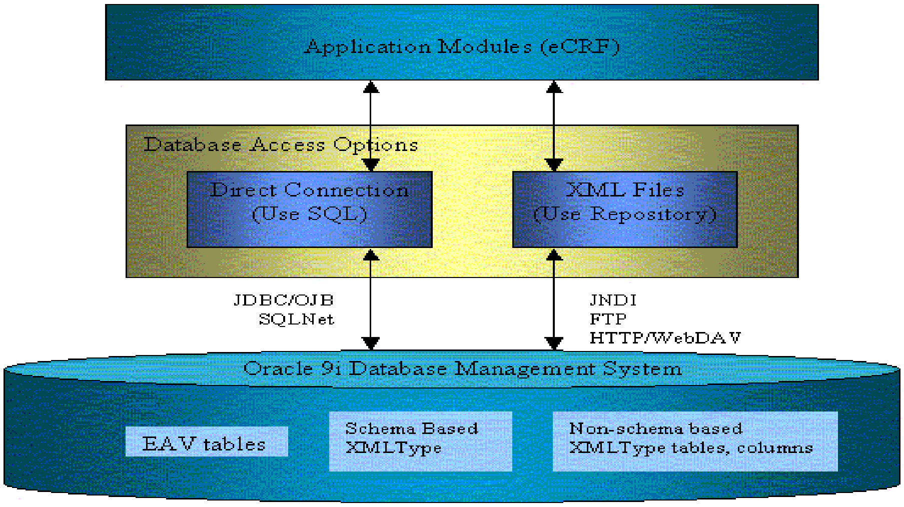
Application Modules (eCRF)
Database Access Options
Direct Connection (Use SQL)
XML Files (Use Repository)
JDBC/OJB
SQLNet
JNDI
FTP
HTTP/WebDAV
Oracle 9i Database Management System
EAV tables
Schema Based XMLType
Non-schema based XMLType tables, columns

Figure 1: Data storage schematic for CSIS

In an "EAV" design, the attributes for an entity are not hardwired into the database as table columns as in a traditional relational database. Rather, they are stored as data, one row for every attribute. This design is often referred to as “vertical design”, or “row modeling”. In addition, metadata describing each data element are stored in a data library, where the data item definitions can be readily created, viewed, and edited by the user. One possible structure for this metadata could be the CDISC ODM. The EAV design makes it possible to accommodate new protocols (with new data items) without the additional programming that would be required in a traditional database design. One needs only to add a description of each new data element to the data library. A good example of clinical information system implementing EAV design is Yale University’s TrialDB.[3]

Alternatively, a XML Schema can be created when an investigator creates a new clinical form. Later, when the form is filled, an instance of this XML Schema can be transported through Oracle XMLDB and stored in the Oracle’s new XMLType tables or columns. Oracle XML DB is a set of utilities in Oracle 9i Release 2 that provides native support for storing and retrieving XML elements from XML documents. It stores information within the Oracle database and represents underlying data “dually”, both as sets of XML elements within XML documents and as cells within relational tables. This structure allows for fine-grained queries on the data contained in the XML document, utilizing the traditional RDBMS tuning mechanisms (e.g. indexes and partitions), while maintaining DOM fidelity for viewing the entire document at once. This structure also lends itself to utilizing the CDISC ODM data format as a metadata definition. In fact, since the ODM structure is already available as an XML schema, we feel that utilizing the Oracle XMLType is more natural choice than EAV, provided that the performance of Oracle XMLDB is acceptable.

Oracle XMLDB Performance Test

In order to test the performance of implementing CDISC ODM in Oracle XMLDB structures, we have simulated the effect of multiple Oracle users performing real world database transactions. A standard clinical form (a psychiatric rating scale form) was implemented in CDISC ODM format. Two storage options were tested: 1) XMLDB unstructured Character Large Object (CLOB) storage and 2) XMLDB schema-based structured storage. In the XMLDB unstructured storage option, a query for element within the document is specified using the XML Path Language, Xpath, such that every time an element is queried the whole document has to be parsed into memory. In the schema-based structured storage option, Oracle parses the document and stores the data in an object-relational structure that was created when the schema was registered with the database. A aggregate mix of database transactions was tested in a "throughput" test, including insertion and retrieval of a complete form, updating a field within that form and returning a list of patients that match a statistical criterion. We used the software "Benchmark Factory for Oracle" from Quest Software to conduct the throughput benchmark test. A summary of the test results is shown in Figure 2. The test results show that the second option, storing clinical data in XML CLOB, results in faster insertions as well as faster full scan queries.

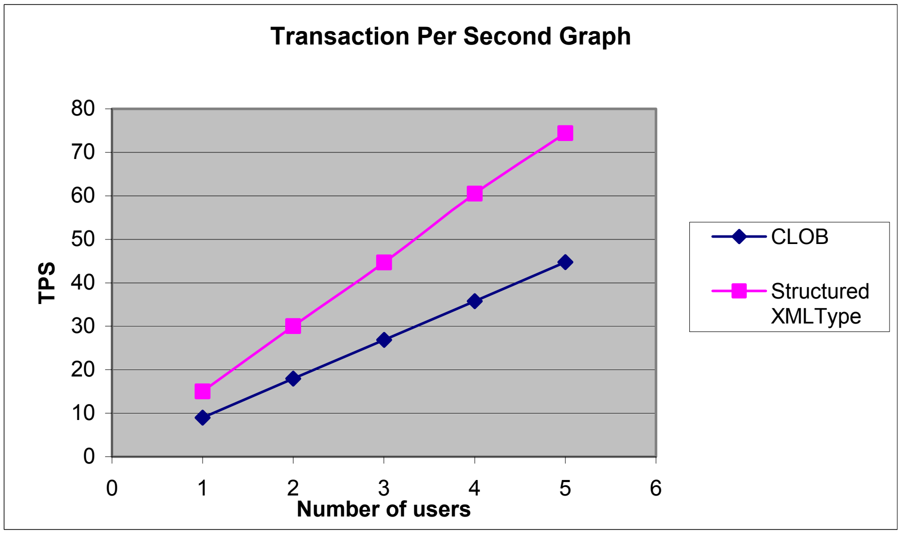

Figure 2 – Oracle XMLDB Test Results. The test was performed on 10,000 records occupying a total of 100 MB of disk space. The test was performed on a 2-GHz dual-processor Dell Pentium 4 server equipped with 12GB RAM running Redhat 9 and Oracle 9i version 9.2.04 respectively.

The results from Figure 2 suggest that structured XML storage outperforms CLOB storage in an aggregate throughput test. However, we have also discovered that CLOB storage outperforms structured XML storage in certain tests, including inserting full XML documents (1.933 seconds versus 17.53 seconds on the same data) and full scan queries that return full documents (16.08 seconds versus 57.76 seconds). The structured tables were not tuned or indexed.

Conclusions

Based on the results from the performance test, the following architectural decisions were made for CIMS. The dynamic clinical form questions and layout will be stored in CLOB format to take advantage of fast storage and retrieval of complete documents. The patient data will be stored using structured XML to facilitate efficient data analysis and reporting. We have concluded that this storage mechanism combined with centralized data dictionary provides the flexibility required and ensures adequate performance for the CIMS project.

References

1. Kush R, "A Multidisciplinary Approach to Data Standards for Clinical Development - Progress Update", Applied Clinical Trials, vol. 11, no. 4 (April 2002), pp. 35--44 (2002).
2. McCray AT and Bodenreider O, "A Conceptual Framework for the Biomedical Domain," in *The symantics of relationships: an interdisciplinary perspective*, Green R, Bean CA, and Myaeng SH, eds., Boston: Kluwer; 2002, pp. 181-198.
3. Grant AM, Delisle E, Perras D, Beteau M, Xhignesse M., "Appreciation of the need for informatics support in applied clinical research", *Proc AMIA Symp* 1997; 857-860.